

Introduction to the Practice of
STATISTICS NINTH
EDITION

Moore / McCabe / Craig

Kapittel 1:

Utforske data — Fordelinger

Lecture Presentation Slides

Macmillan Learning © 2017

Kapittel 1

Utforske data — Fordelinger

Introduksjon

1.1 Data

1.2 Presentere fordelinger med grafer

1.3 Beskrive fordelinger med tall

1.4 Tetthetskurver og normalfordelingen

1.1 Data

- Viktige begreper: Enheter, variable og verdier
- Kategoriske og kvantitative variable
- Viktige aspekter ved et datasett

Viktige begreper

- ✓ **Enheter** er objektene som blir beskrevet av et datasett. Enheter kan være individer i en studie (dyr, mennesker osv.), enheter i et eksperiment eller andre objekter.
- ✓ En **variabel** er en spesifikk karakteristikk av hver enhet.
- ✓ Variablene angis med **verdier**, som varierer mellom enhetene.
- ✓ En **merkelapp** er en spesiell variabel som brukes i noen datasett for å skille mellom ulike enheter.

Kategoriske og kvantitative variable

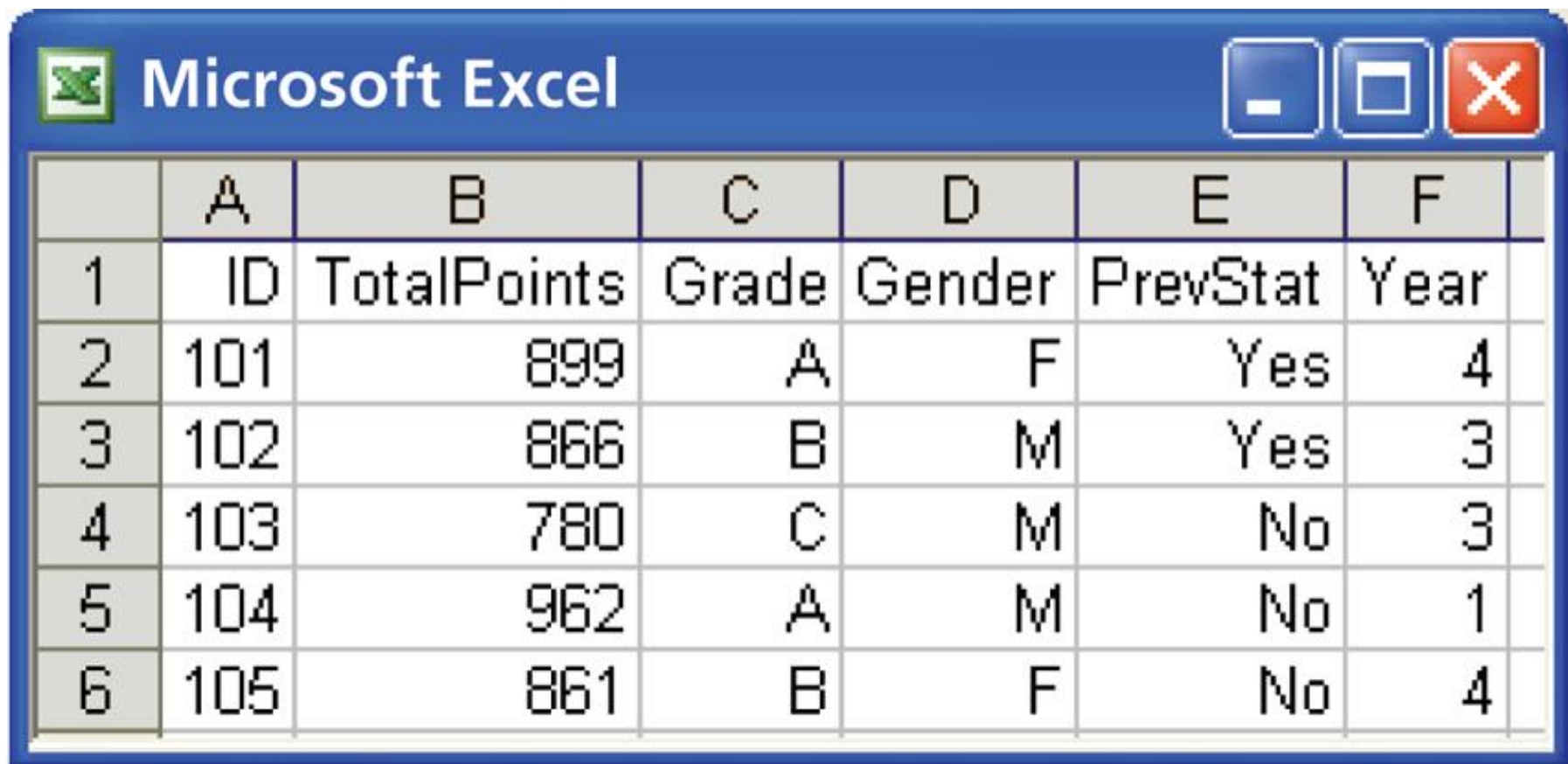
- ❑ En **kategorisk** variabel plasserer hver enhet i en av flere grupper, eller kategorier.
- ❑ En **kvantitativ** variabel tar numeriske verdier, som gjør det meningsfullt å foreta aritmetiske operasjoner som for eksempel å summere og beregne gjennomsnitt.

Eksempel på datasett

6

Data på studenter som tar et innføringsemne i statistikk

- Enheter?
- Variable? Kategoriske eller kvantitative?



The image shows a screenshot of a Microsoft Excel spreadsheet. The window title is "Microsoft Excel". The spreadsheet contains the following data:

	A	B	C	D	E	F
1	ID	TotalPoints	Grade	Gender	PrevStat	Year
2	101	899	A	F	Yes	4
3	102	866	B	M	Yes	3
4	103	780	C	M	No	3
5	104	962	A	M	No	1
6	105	861	B	F	No	4

Viktige aspekter ved et datasett

7

Med alle datasett følger det viktig bakgrunnsinformasjon. I en statistisk studie **bør man alltid stille seg følgende spørsmål:**

- **Hvem?** Hvilke **enheter** beskrives av dataene? **Hvor mange** enheter er det i datasettet?
- **Hva?** Hvor mange **variabler** har datasettet? Hva er de **eksakte definisjonene** av disse variablene? Hva er måleenhetene for hver av de kvantitative variablene?
- **Hvorfor? Hvilket formål** har dataene? Inneholder dataene den nødvendige informasjonen for å besvare spørsmålene man er interesserte i?

1 Kahoot-spørsmål

1.2 Presentere fordelinger med grafer

- Eksplorativ dataanalyse
- Grafer for kategoriske variable
 - Stolpediagrammer
 - Kakediagrammer
- Grafer for kvantitative variabler
 - Histogrammer
 - Stilk- og blad-plott

Utforske data

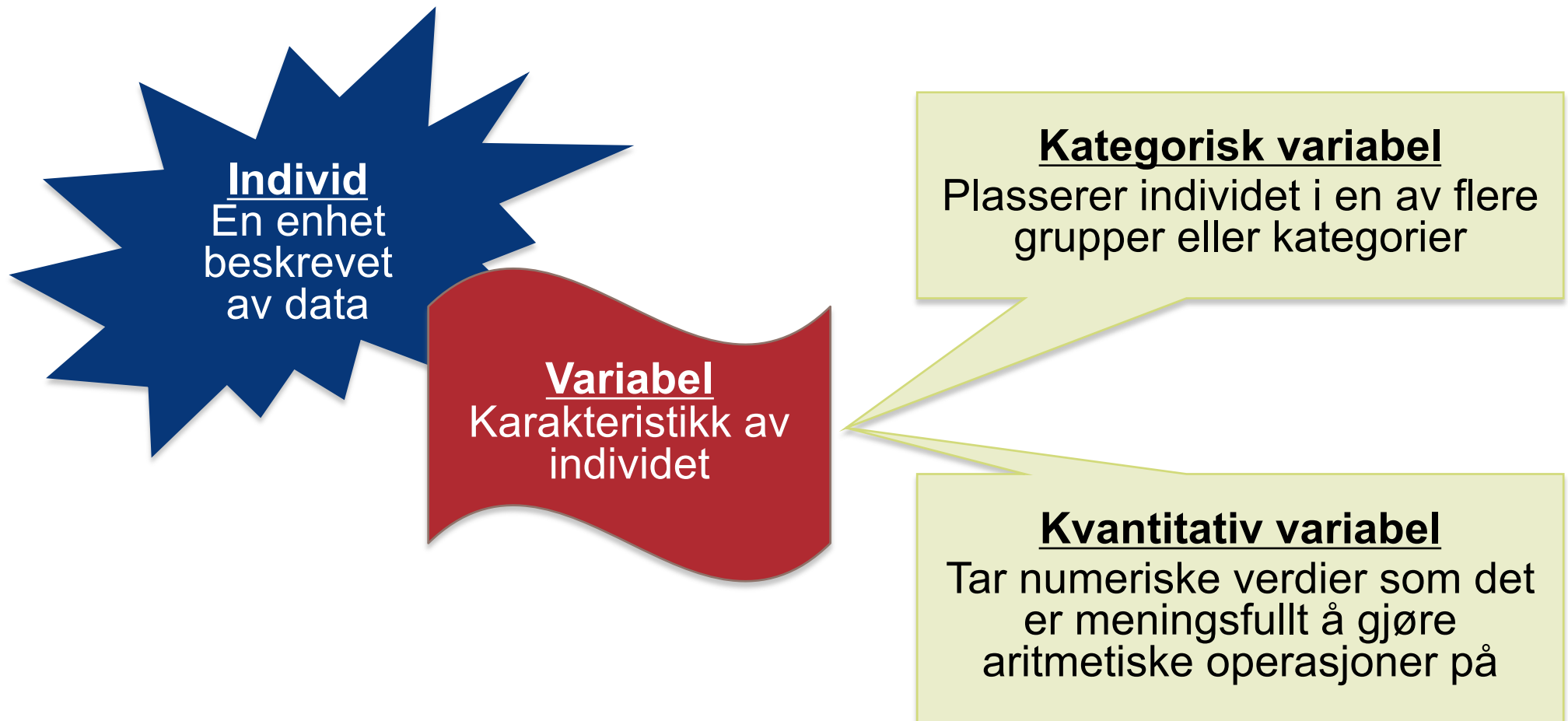
- Begynn med å utforske hver variabel for seg. Fortsett med å se på sammenhenger mellom variablene.
- Begynn med en eller flere grafer. Fortsett så med å se på numeriske oppsummeringer av spesifikke aspekter ved dataene.

Variabler

11

Et nytt datasett lages ved å først bestemme seg for **enheter** det er ønskelig å studere.

For hver enhet samles det inn informasjon om karakteristikk som vi kaller **variabel**.



Fordelingen til en variabel

12

For å undersøke en enkelt variabel viser vi **fordelingen** til variabelen grafisk.

- Fordelingen til en variabel forteller oss hvilke verdier den tar og hvor ofte den tar disse verdiene.
- Fordelinger kan vises ved å bruke mange forskjellige grafiske verktøy. Valg av type graf avhenger av hva slags variabel det er snakk om.

Kategorisk variabel

Kakediagram
Stolpediagram

Kvantitativ variabel

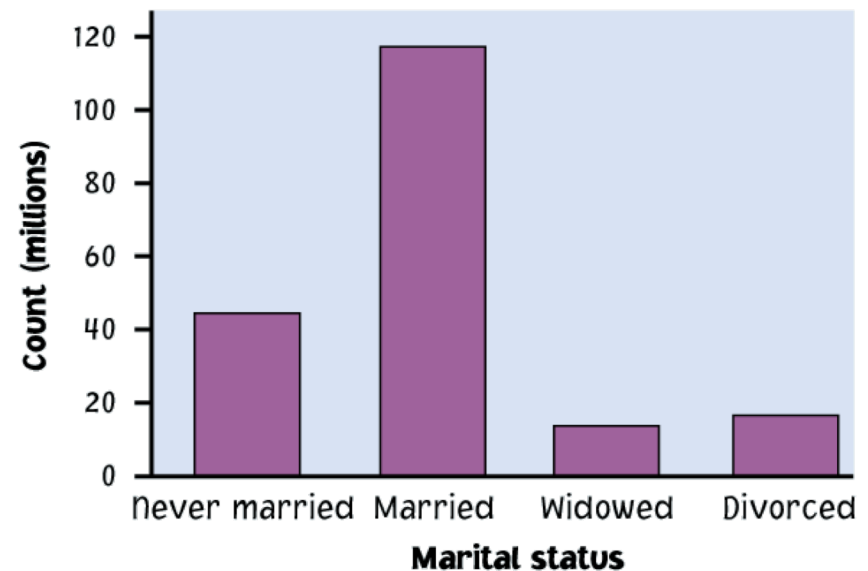
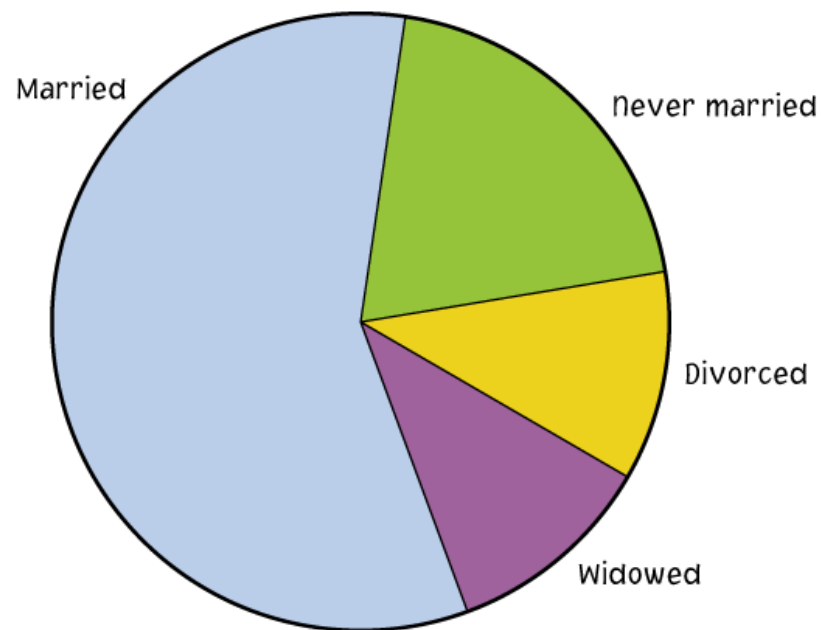
Histogram
Stilk- og blad-plott

Kategoriske Variable

13

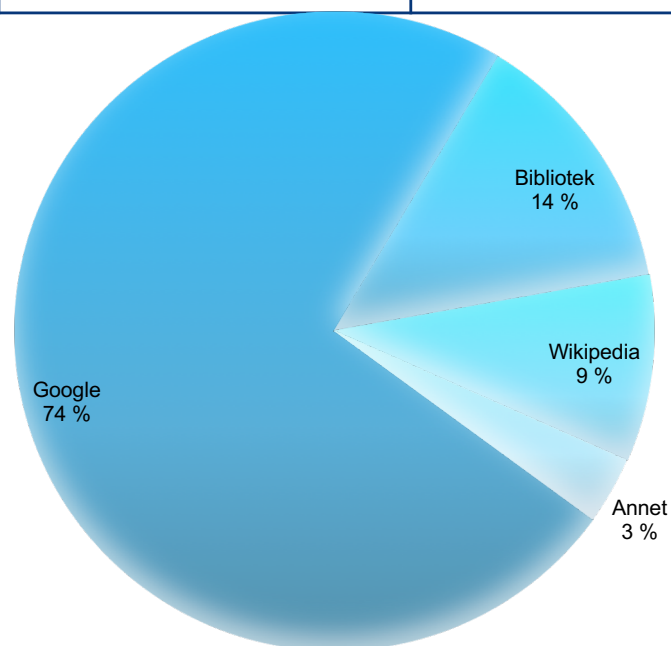
Fordelingen til en kategorisk variabel viser kategoriene og gir **antallet** eller **prosentandel** av individer som faller i hver av kategoriene.

- **Kakediagrammer** viser fordelingen til en kategorisk variabel som en “kake” der kakestykkenes størrelser tilsvarer prosentandeler for kategoriene, med hele kaken tilsvarende 100%.
- **Stolpediagrammer** representerer kategorier som stolper med høyder som tilsvarer antall eller prosentandeler for kategoriene.

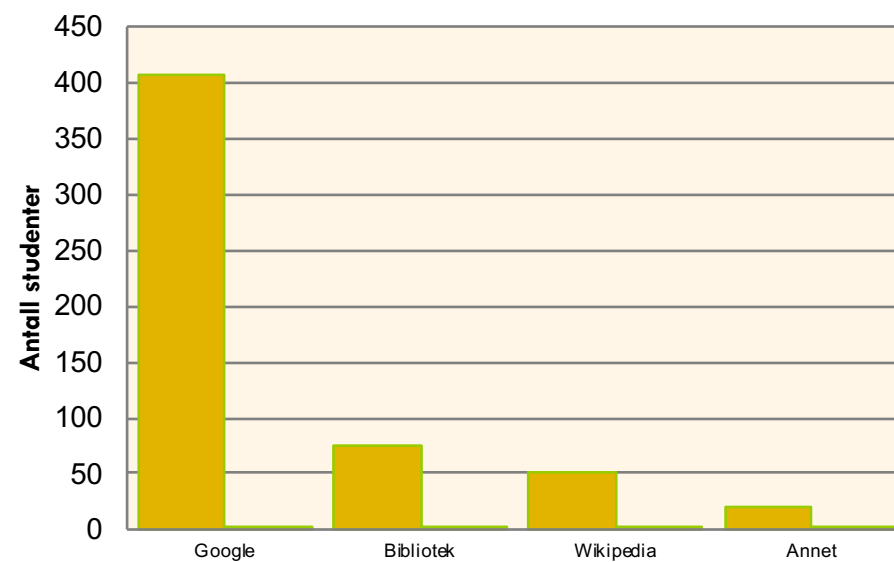


Kake- og stolpe-diagrammer

Online forskningskilder foretrukket av studenter	Antall	Prosentandel av totalen
Google eller Google Scholar	406	73.6%
Biblioteks-database eller webside	75	13.6%
Wikipedia eller online leksikon	52	9.4%
Annet	19	3.4%
Totalt	552	100.0%



Online forskningskilder



Fordelingen til en kvantitativ variabel forteller oss hvilke verdier variabelen tar og hvor ofte den tar de verdiene.

- **Stilk- og blad-plott** deler hver observasjon i en stamme og et blad som blir plottet for å vise fordelingen og samtidig beholde de opprinnelige observerte verdiene til variabelen- *nyttig når datamengden er liten*
- **Histogrammer** viser fordelingen til en kvantitativ variabel ved å bruke stolper for intervaller av verdier. Høyden til en stolpe representerer antall individer som har verdier som faller innenfor det tilsvarende intervallet- *nyttig for store datamengder*

Stilk- og blad-plott 1

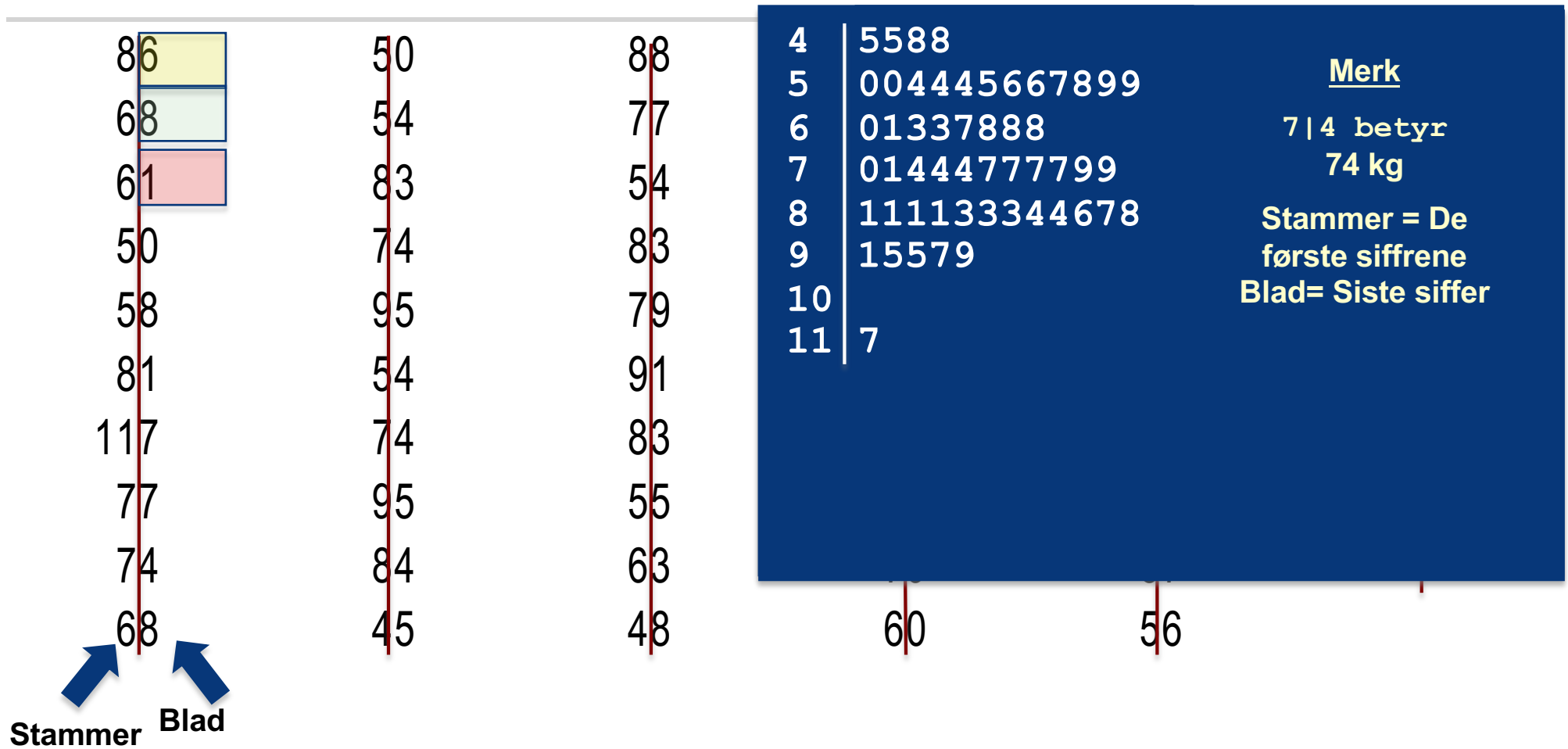
Oppskrift for å lage stilk- og blad-plott:

- Del hvert observert tall opp i en **stamme** (alle bortsett fra det siste sifferet) og et **blad** (det siste sifferet i tallet).
- Skriv stammene i en vertikal kolonne (i første omgang bare ett eksemplar av hver stamme); tegn så en vertikal linje til høyre for stammene.
- Skriv hvert blad i raden til høyre for stammen sin; sorter gjerne bladene.

Hvis de observerte tallene har mange sifre, kan det være hensiktsmessig å trimme tallene ved å fjerne det/de siste sifferet/sifrene før man lager plottet.

Stilk- og blad-plott 2

Eksempel: Vektdata — Introduksjonsemne i statistikk



Stilk- og blad-plott 3

Hvis det er veldig få stammer (dataene dekker bare et lite intervall av verdier), ønsker vi kanskje å skrive opp flere stammer ved å **dele opp** de opprinnelige stammene.

Eksempel: Hvis alle dataverdiene ligger mellom 150 og 179, ønsker vi kanskje å velge følgende stammer:

15
15
16
16
17
17

Blader 0–4 vil settes ved den øverste av de tilhørende stammene, og blader 5–9 settes ved den nederste av de tilhørende stammene

Stilk- og blad-plott 4

Eksempel på å dele opp stammene (opptak av kalsium for kontroll-gruppe og en gruppe som inntar SCF-løselig mais-fiber):

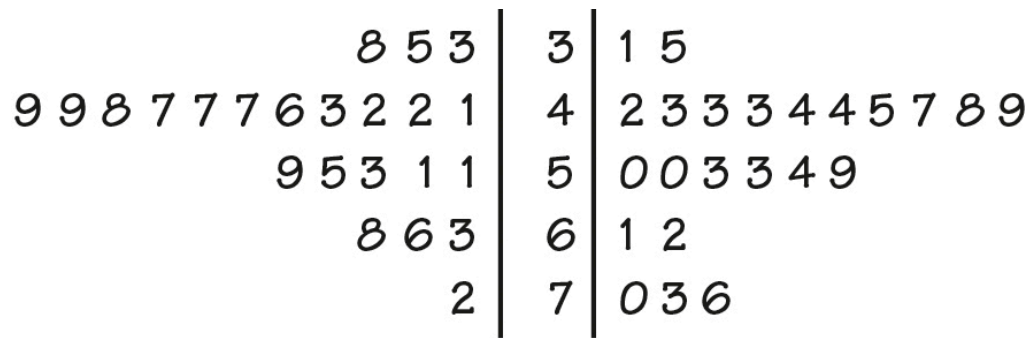


Figure 1.5

Moore/McCabe/Craig, *Introduction to the Practice of Statistics*, 9e,
© 2017 W. H. Freeman and Company

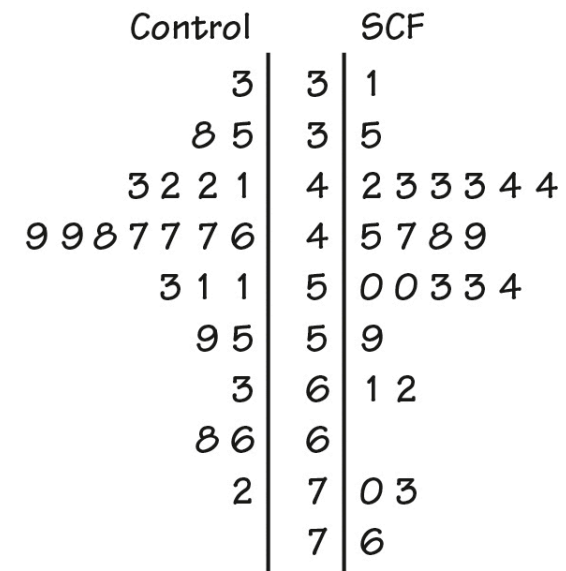


Figure 1.6

Moore/McCabe/Craig, *Introduction to the Practice of Statistics*, 9e, © 2017 W. H. Freeman and Company

For kvantitative datasett som er store:

- Del de mulige verdiene inn i **klasser** eller intervaller som er like brede.
- Tell hvor mange observasjoner som faller i hvert intervall. I stedet for antall, kan man bruke prosentandeler.
- Tegn et bilde som representerer fordelingen- høyden på hver stolpe tilsvarer antall (eller prosentandel) observasjoner i det tilhørende intervallet.

Histogrammer: Eksempel

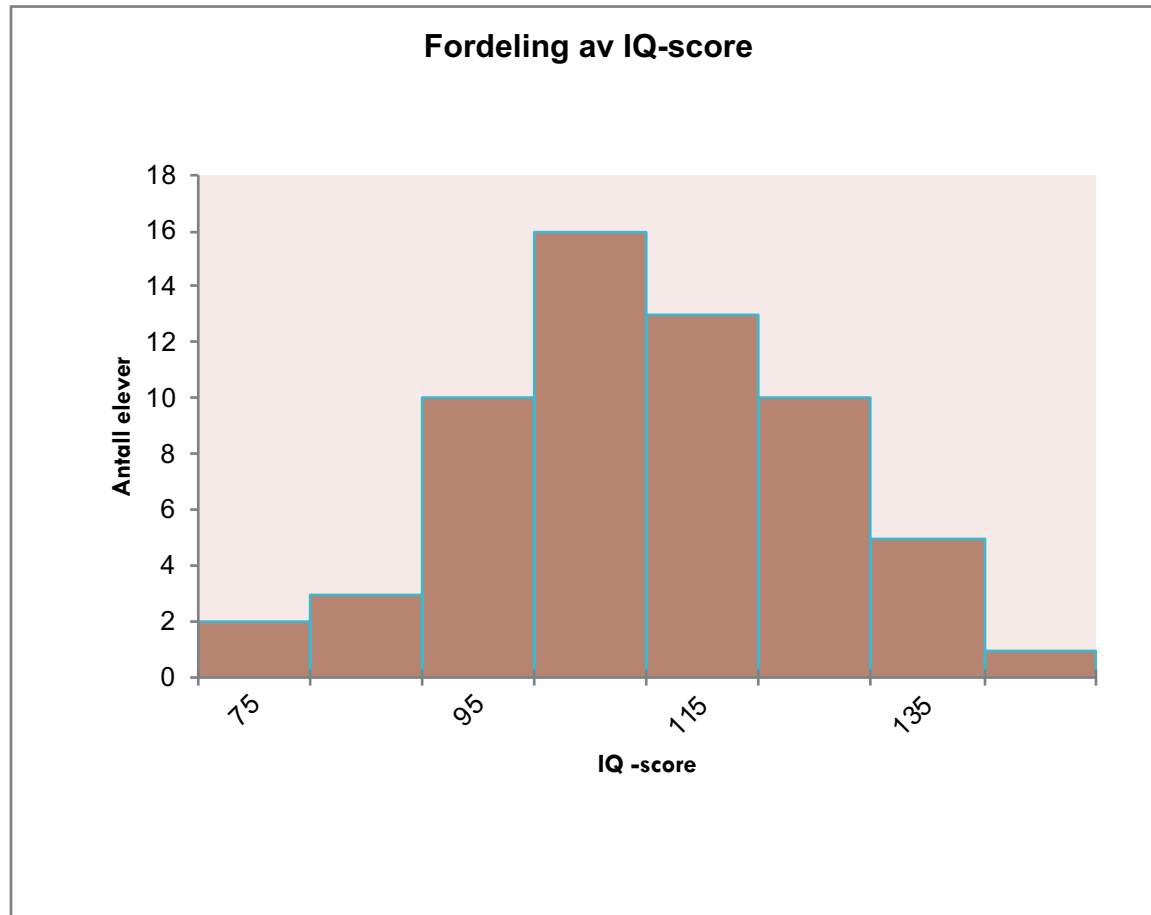
IQ-score for 60 5.-klassinger

145	139	126	122	125	130	96	110	118	118
101	142	134	124	112	109	134	113	81	113
123	94	100	136	109	131	117	110	127	124
106	124	115	133	116	102	127	117	109	137
117	90	103	114	139	101	122	105	97	89
102	108	110	128	114	112	114	102	82	101

Klasse	Antall
$75 \leq \text{IQ-score} < 85$	2
$85 \leq \text{IQ-score} < 95$	3
$95 \leq \text{IQ-score} < 105$	10
$105 \leq \text{IQ-score} < 115$	16
$115 \leq \text{IQ-score} < 125$	13
$125 \leq \text{IQ-score} < 135$	10
$135 \leq \text{IQ-score} < 145$	5
$145 \leq \text{IQ-score} < 155$	1

Histogrammer: Eksempel

IQ-score for 60 5.-klassinger



Klasse	Antall
$75 \leq \text{IQ-score} < 85$	2
$85 \leq \text{IQ-score} < 95$	3
$95 \leq \text{IQ-score} < 105$	10
$105 \leq \text{IQ-score} < 115$	16
$115 \leq \text{IQ-score} < 125$	13
$125 \leq \text{IQ-score} < 135$	10
$135 \leq \text{IQ-score} < 145$	5
$145 \leq \text{IQ-score} < 155$	1

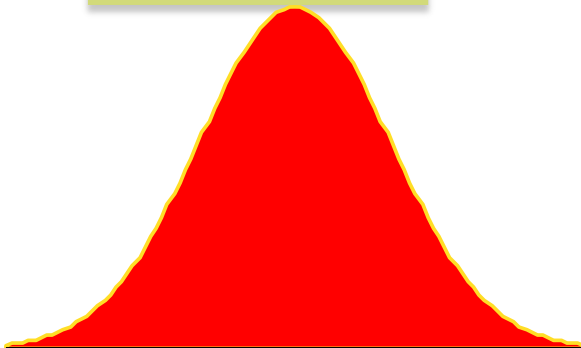
- Ved enhver grafisk fremstilling av data, se etter det **overordnede mønsteret** og etter fremtredende **avvik** fra det mønsteret.
- Du kan beskrive det overordnede mønsteret ved **form**, **senterpunkt**, og **spredning**.
- En viktig type avvik er en **uteligger** ('outlier' på engelsk), et individ som faller utenfor det overordnede mønsteret.

Beskrive formen på en fordeling

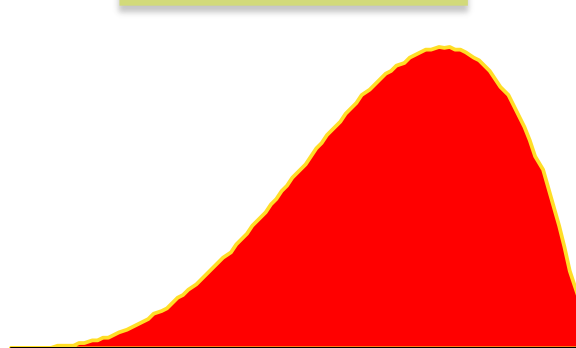
24

- Hvor mange topper (også kalt moder) har fordelingen? En fordeling med bare en topp kalles **unimodal**, en fordeling med to topper **bimodal**.
- En fordeling er **symmetrisk** hvis høyre og venstre siden av grafen er tilnærmet speilbilder av hverandre.
- Den er **venstre-skjev** hvis venstre side av grafen er mye lengre enn høyre side.
- En fordeling er **høyre-skjev** hvis høyre side av grafen (som inneholder den halvparten av observasjonene som har høye verdier) er mye lengre enn venstre side av grafen.

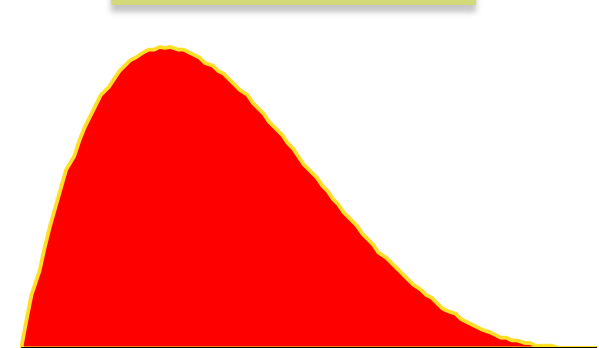
Symmetrisk



Venstre-skjev



Høyre-skjev

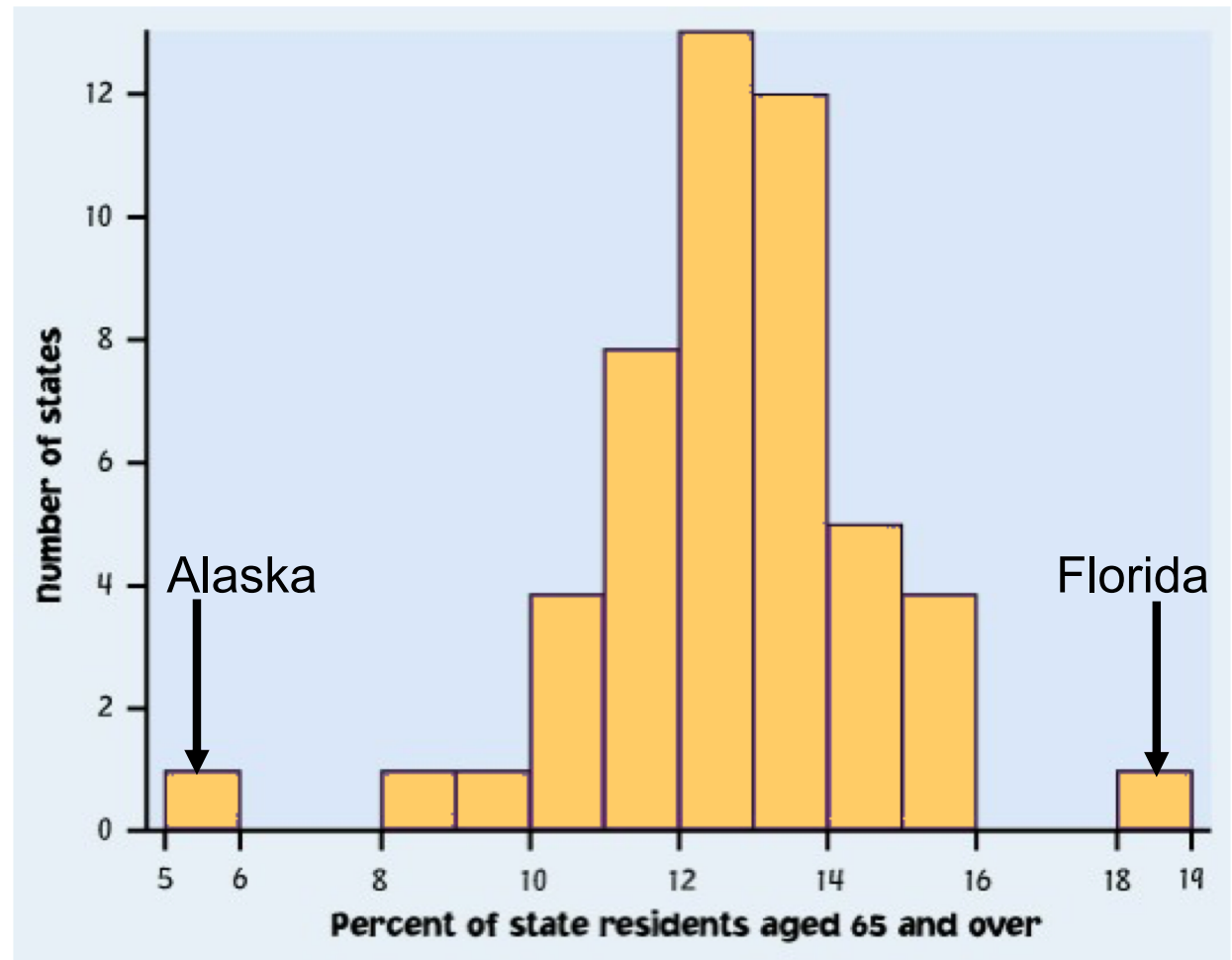


Uteliggere

En viktig type avvik er en **uteligger**. Uteliggere er observasjoner som ligger utenfor det overordnede mønsteret. Se alltid etter uteliggere og prøv å finne en forklaring for dem.

Det overordnede mønsteret er en topp, relativt symmetrisk - med unntak av to stater som tydelig ikke hører til hoved-mønsteret. Alaska og Florida har uvanlig henholdsvis små og store prosentandeler av eldre innbyggere i deres befolkninger.

Et stort gap i fordelingen er et typisk tegn på en uteligger.



Et **tidsrekke-plott** viser oppførsel over tid.

- Tid er alltid langs den horisontale aksene, og variabelen som måles er langs den vertikale aksene.
- Se etter et overordnet mønster (trend) og avvik fra denne trenden. Å knytte datapunktene sammen via linjer kan understreke denne trenden.
- Se etter mønstre som repeteres i kjente, regulære intervaller (sesong-variasjoner).

Tidsrekkeplott 2

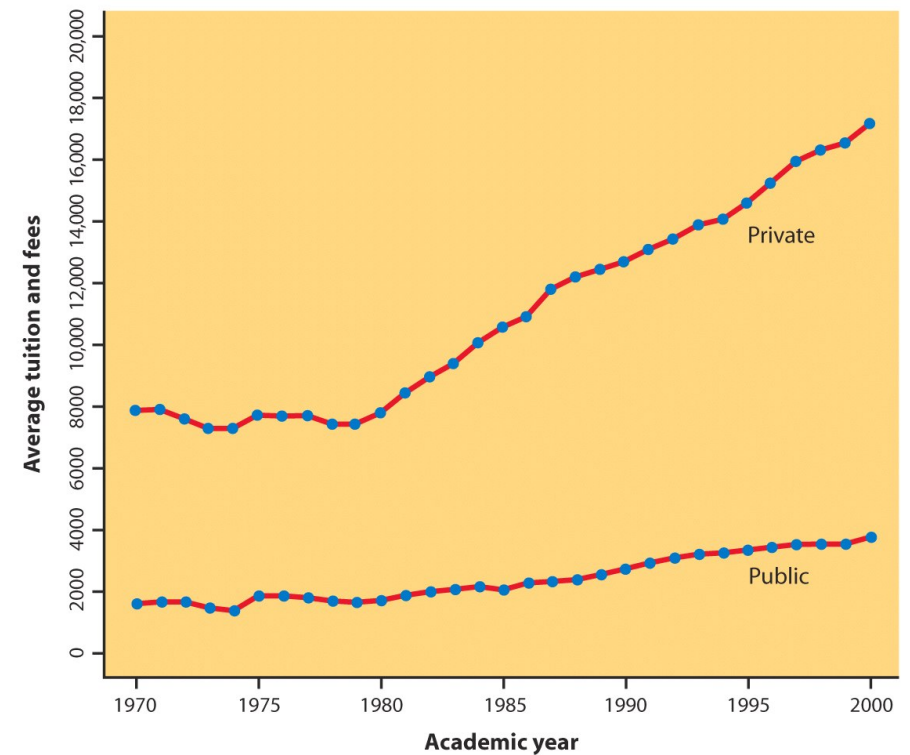
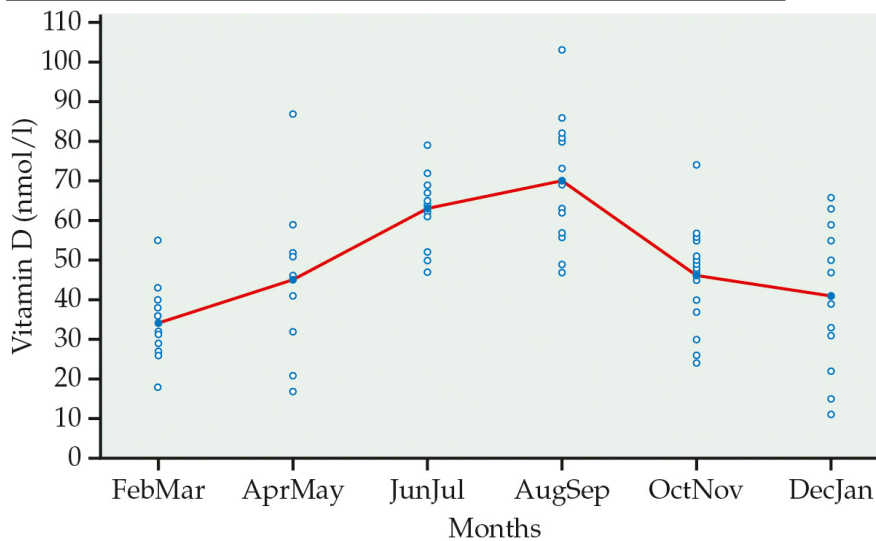
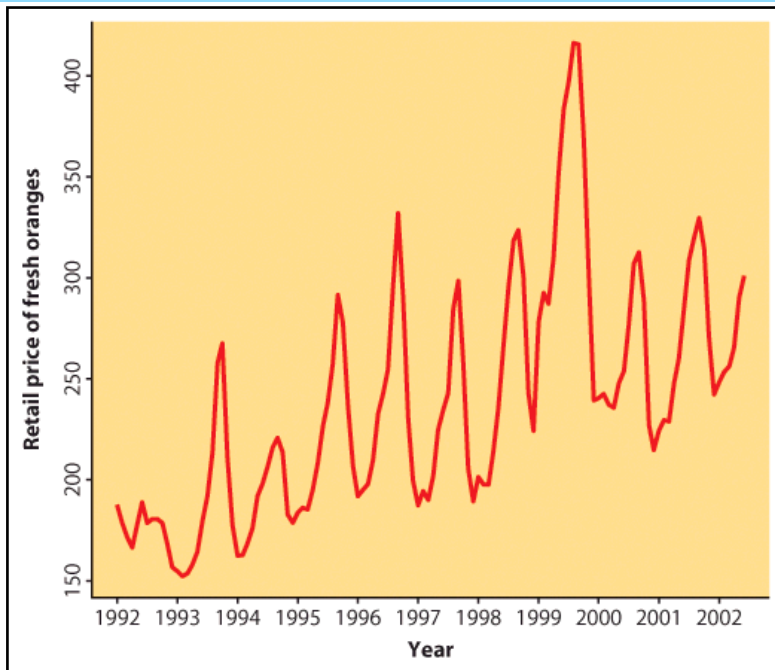


Figure 1.12
Moore/McCabe/Craig, *Introduction to the Practice of Statistics*, 9e, © 2017 W. H. Freeman and Company

2 Kahoot-spørsmål

1.3 Beskrive fordelinger med tall

- Mål på senter: gjennomsnitt
- Mål på senter: median
- Mål på spredning: kvartiler
- Fem-punkts-oppsummering og boksplott
- Mål på spredning: standardavvik
- Å velge mellom oppsummerende mål
- Forandre måleenhet

Mål på senter: Gjennomsnittet

Det mest vanlige målet på senter er **gjennomsnittet**.

For å finne **gjennomsnittet** \bar{x} (uttales “x-bar”) av en mengde observasjoner, summer verdiene deres og del på antall observasjoner. Hvis de n observasjonene er $x_1, x_2, x_3, \dots, x_n$, er gjennomsnittet

$$\bar{x} = \frac{\text{sum of observasjoner}}{n} = \frac{x_1 + x_2 + \dots + x_n}{n}$$

I en mer kompakt, matematisk notasjon:

$$\bar{x} = \frac{1}{n} \sum x_i$$

Mål på senter: Medianen

31

Fordi gjennomsnittet er sensitivt for ekstreme observasjoner, er det ikke et **robust** mål på senter.

Et vanlig og mer robust mål på senter er **medianen**.

Medianen M er midtpunktet i fordelingen, tallet som er slik at halvparten av observasjonene er mindre og den andre halvparten er større. *Dette påvirkes mindre av ekstreme observasjoner enn gjennomsnittet, fordi det ikke direkte bruker verdien av hver enkelt observasjon.*

Medianen i en fordeling finnes på følgende måte:

1. Ordne alle observasjonene i stigende rekkefølge.
2. Hvis antallet observasjoner n er odde, er medianen M den midterste observasjonen i den ordnede listen.
3. Hvis antallet observasjoner n er like, er medianen M gjennomsnittet av de to midterste observasjonene i den ordnede listen.

Mål på senter: Eksempel

Bruk dataene under til å beregne gjennomsnittet og medianen til tiden det tar å etablere en bedrift (i dager) i 24 tilfeldig valgte land.

16	4	5	6	5	7	12	19	10	2	25	19
38	5	24	8	6	5	53	32	13	49	11	17

$$\bar{x} = \frac{16 + 4 + 5 + 6 + \dots + 11 + 17}{24} = 16.292 \text{ dager}$$

0	2455556678
1	0 1 236799
2	45
3	28
4	9
5	3

Merk: 4|9
representerer et
land der det tar 49
dager å starte en
bedrift.

$$M = \frac{11 + 12}{2} = 11.5 \text{ dager}$$

Sammenligne gjennomsnitt og median

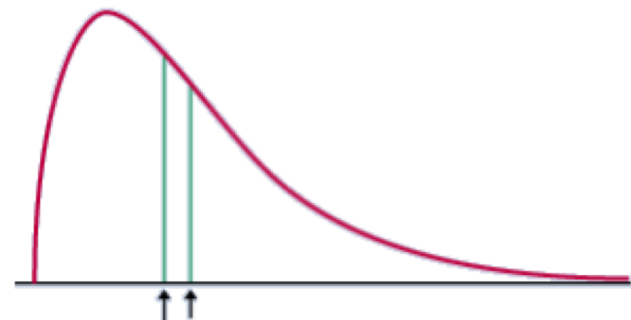
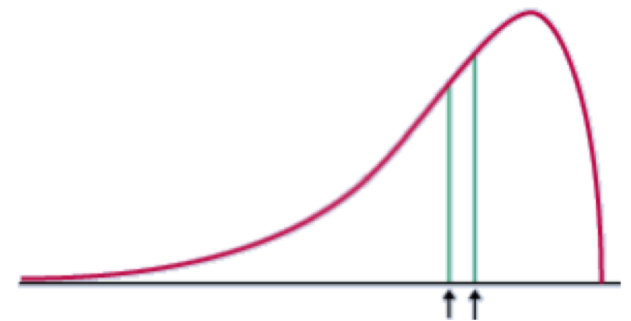
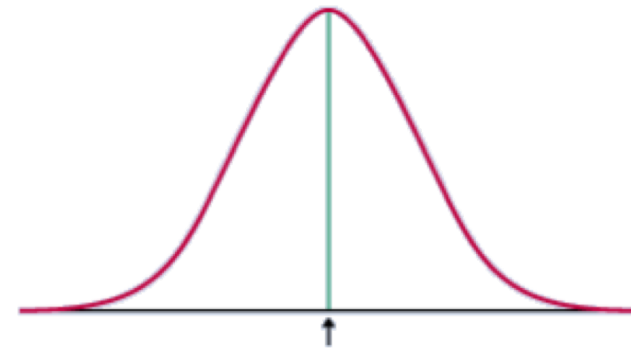
33

Gjennomsnittet og medianen måler senter på forskjellige måter, og begge er nyttige.

Gjennomsnittet og medianen til en tilnærmet **symmetrisk** fordeling er veldig like.

Hvis fordelingen er eksakt **symmetrisk** er gjennomsnittet og medianen eksakt like.

I en **skjev** fordeling er gjennomsnittet lenger ut i halen enn medianen.



- Å måle senter alene kan være villedende.
- For at en numerisk beskrivelse av data skal være nyttig må man ha både mål på senter og mål på spredning.

Å beregne kvartilene

- Ordne observasjonene i stigende rekkefølge og finn **medianen M** .
- Den **første kvartilen Q_1** er medianen til observasjonene som befinner seg *venstre* for medianen i den ordnede listen. Dvs 25% av dataene er mindre enn Q_1 , og 25% av dataene er mellom Q_1 og M .
- Den **tredje kvartilen Q_3** er medianen til observasjonene som befinner seg til *høyre* for medianen i den ordnede listen. Dvs 25% av dataene er større enn Q_3 , og 25% av dataene er mellom M og Q_3 .
- Sammen med medianen deler altså kvartilene dataene inn i fire like deler: 25% av dataene er i hver del.

Kvartilene er ikke veldig sensitive for ekstreme observasjoner.

Mer generelt: Den p -te **percentilen** er den verdien som er slik at p % av dataene er mindre enn den.

Fem-punkts-oppsumming

Medianen og kvartilene er mål på senter og spredning, men forteller oss lite om halene i fordelingen. Minimum- og maksimums-verdiene sier noe om halene, men forteller oss lite om helheten i fordelingen.

For å få en rask oppsummering av både senter og spredning kan man kombinere alle disse fem tallene.

Fem-punkts-oppsummingen til en fordeling består av den minste observasjonen, den første kvartilen, medianen, den tredje kvartilen, og den største observasjonen, presentert i stigende rekkefølge.

Minimum Q_1 M Q_3 *Maksimum*

Mistenkte uteliggere: $1.5 \times IQR$ -regelen

36

Mål på spredning: **Interkvartil-avstand (IQR)**, definert som $IQR = Q_3 - Q_1$. Ikke veldig sensitivt for ekstreme observasjoner.

I tillegg til å være et mål på spredning brukes IQR som del av en tommelfingerregel for å identifisere uteliggere.

$1.5 \times IQR$ -regelen for uteliggere

En observasjon er en uteligger dersom den faller mer enn $1.5 \times IQR$ over den tredje kvartilen eller under den første kvartilen.

I bedriftstart-dataene, $Q_1 = 5.5$ dager, $Q_3 = 21.5$ dager, og dermed $IQR = 16$ dager.

For disse dataene, $1.5 \times IQR = 1.5(16) = 24$

$$Q_1 - 1.5 \times IQR = 5.5 - 24 = -18.5$$

$$Q_3 + 1.5 \times IQR = 21.5 + 24 = 45.5$$

Ethvert tilfelle der det tar kortere enn -18.5 dager eller lenger enn 45.5 dager å starte en bedrift anses som uteligger.

0	245556678
1	01236799
2	45
3	28
4	9
5	3

1 Kahoot-spørsmål

Medianen og kvartilene deler fordelingen i kvartdeler. Dette gir oss en ny metode for å grafisk fremstille kvantitative data, som vi kaller et **boksplott**.

Hvordan lage et boxplott

- Tegn og en akse som inkluderer hele verdiområdet til fordelingen.
- Tegn en boks langs aksene som har en ende i Q_1 og den andre enden i Q_3 .
- Merk av medianen M inne i boksen.
- Tegn linjer (værhår) fra boksen ut til minimum- og maksimum-verdiene som ikke er uteliggere.
- Uteliggere markeres som punkter

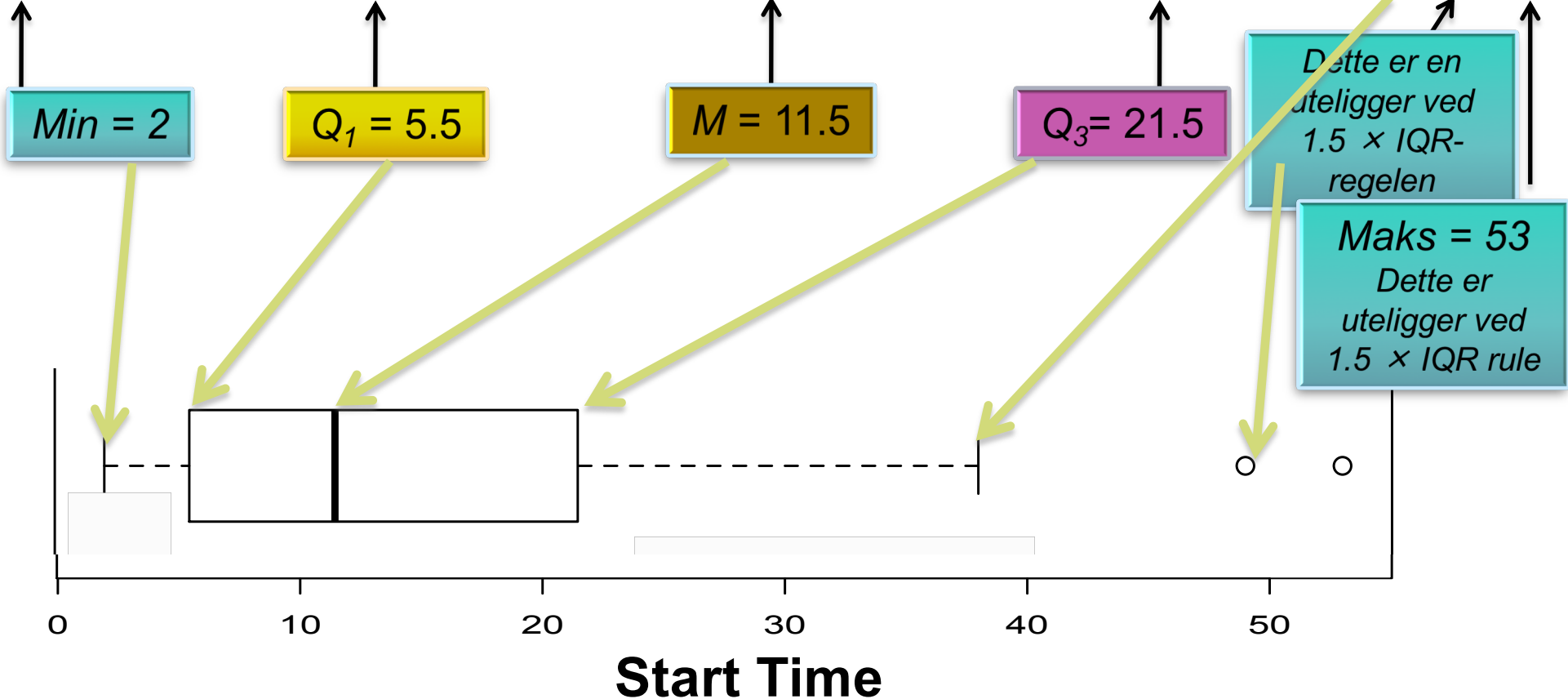
Boxplott 2

Ser igjen på data på hvor lang tid det tar å starte en bedrift i ulike land. Konstruer et boksplott.

16	4	5	6	5	7	12	19	10	2	25	19
38	5	24	8	6	5	53	32	13	49	11	17

Sorter dataene

2	4	5	5	5	5	6	6	7	8	10	11	12	13	16	17	19	19	24	25	32	38	49	53
---	---	---	---	---	---	---	---	---	---	----	----	----	----	----	----	----	----	----	----	----	----	----	----



Mål på spredning: Standardavviket

Det mest vanlige målet på spredning ser på hvor langt hver observasjon er fra gjennomsnittet. Dette målet kalles **standardavviket**.

Standardavviket s_x er et slags mål på gjennomsnittlig avstand til observasjonene fra deres gjennomsnitt. Det beregnes ved å finne et gjennomsnitt av de kvadrerte avstandene og deretter ta kvadratroten. Den gjennomsnittlige kvadrerte avstanden kalles **variansen**

$$\text{varians} = s_x^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \dots + (x_n - \bar{x})^2}{n - 1} = \frac{1}{n - 1} \times \sum (x_i - \bar{x})^2$$

$$\text{standardavvik} = s_x = \sqrt{\frac{1}{n - 1} \times \sum (x_i - \bar{x})^2}$$

Mål på spredning: Standardavviket

- Negative avstander kansellerer positive
- Summen av ikke kvadrerte avvik fra gjennomsnittet blir alltid null:

$$(x_1 - \bar{x}) + (x_2 - \bar{x}) + \dots + (x_n - \bar{x}) = 0$$

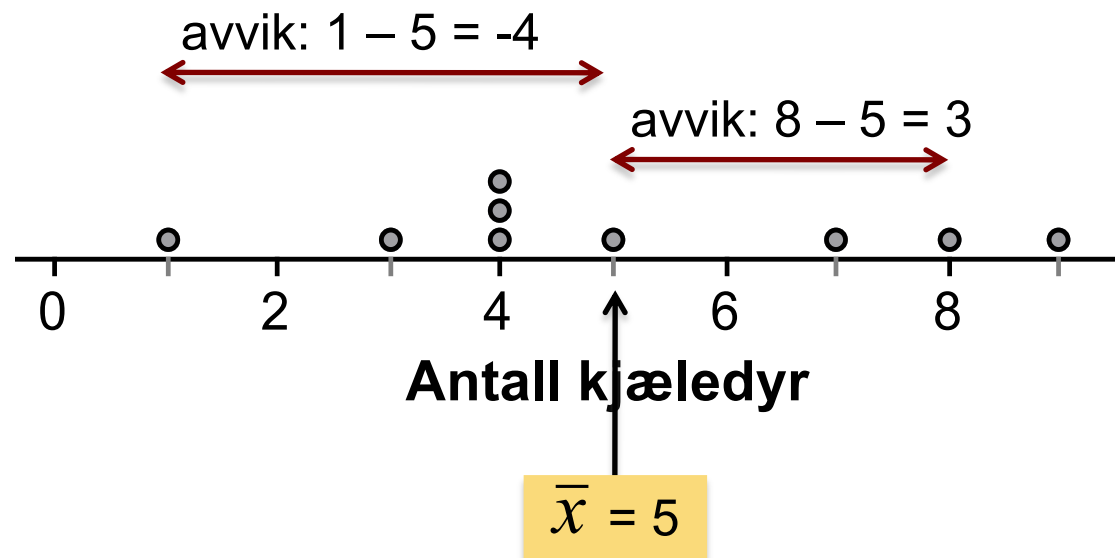
➤ Dette er på grunn av hvordan gjennomsnittet er beregnet

- Å summere de *kvadrerte* avstandene gjør at observasjoner like langt fra hverandre i begge retninger bidrar like mye til spredningsmålet

Beregne standardavviket 1

Eksempel: Vi ser på følgende data på antall kjæledyr ni forskjellige barn har.

1. Beregn gjennomsnittet.
2. Beregn hvert *avvik*.
$$\text{avvik} = \text{observasjon} - \text{gjennomsnitt}$$



Beregne standardavviket 2

3. Kvadrer hvert avvik.
4. Finn “gjennomsnittlig” kvadrert avvik ved å beregne summen av de kvadrerte avvikene delt på $(n - 1)$. Dette kalles **variansen**.
5. Beregn kvadratroten av variansen. Dette er **standardavviket**.

x_i	$(x_i - \text{gj.snitt})$	$(x_i - \text{gj.snitt})^2$
1	$1 - 5 = -4$	$(-4)^2 = 16$
3	$3 - 5 = -2$	$(-2)^2 = 4$
4	$4 - 5 = -1$	$(-1)^2 = 1$
4	$4 - 5 = -1$	$(-1)^2 = 1$
4	$4 - 5 = -1$	$(-1)^2 = 1$
5	$5 - 5 = 0$	$(0)^2 = 0$
7	$7 - 5 = 2$	$(2)^2 = 4$
8	$8 - 5 = 3$	$(3)^2 = 9$
9	$9 - 5 = 4$	$(4)^2 = 16$
	Sum = 0	Sum = 52

“Gjennomsnittlig” kvadrert avvik = $52/(9 - 1) = 6.5$. Dette er **variansen**.

Standardavvik = kvadratrot av varians = $\sqrt{6.5} = 2.55$

Egenskaper til standardavviket

- s måler spredningen rundt gjennomsnittet og bør bare brukes når gjennomsnittet er et passende mål på senter.
- $s = 0$ bare hvis alle observasjoner har samme verdi og det ikke er noen spredning. Ellers er $s > 0$.
- s er *sensitiv* for uteliggere og andre ekstreme observasjoner.
- s har samme måleenhet som de opprinnelige observasjonene.

Velge mål på senter spredning

Vi har nå valget mellom to sett av mål på senter og spredning:

- ✓ Gjennomsnitt og standardavvik
- ✓ Median og inter-kvartil avstand (IQR)

Velge sett av mål på senter og spredning

Medianen og IQR er vanligvis bedre enn gjennomsnitt og standardavvik for å beskrive en skjev fordeling eller en fordeling med sterke uteliggere.

Bruk gjennomsnitt og standardavvik bare for rimelig symmetriske fordelinger som ikke har sterke uteliggere.

NB: Numeriske oppsummeringer beskriver ikke fullt ut formen til en fordeling. *PLOTT ALLTID DATAENE DINE!*

Variabler kan måles med forskjellige måleenheter. Veldig ofte er en måleenhet en **lineær transformasjon** av en annen måleenhet:

$$x_{\text{new}} = a + bx.$$

Eksempel: Temperatur x målt i °F, da er temperaturen målt i °C:

$$x_{\text{new}} = -17.92 + 0.56x, \text{ dvs: } a=-17.92 \text{ og } b=0.56$$

Lineære transformasjoner forandrer ikke formen til en fordeling (skjevhet, symmetri, multimodalitet/flertoppethet). Men de forandrer målene på senter og spredning.

- Å multiplisere hver observasjon med et positivt tall b multipliserer både mål på senter (gjennomsnitt, median) og spredning (IQR, s) med b .
- Å legge tallet a (positivt eller negativt) til hver observasjon legger a også til mål på senter og til kvartiler, men det endrer ikke mål på spredning (IQR, s).

2 Kahoot-spørsmål

1.4 Tetthetskurver og normalfordelinger

- Tetthetskurver
- Mål på senter og spredning for tetthetskurver
- Normalfordelinger
- 68-95-99.7-regelen
- Standardisere observasjoner
- Bruk av standard normalfordelings-tabellen
- Inverse normal-beregninger
- Normalfordelingsplott

Vi har nå en verktøykasse med grafiske og numeriske verktøy for å utforske og beskrive fordelinger til en enkelt kvantitativ variabel.

Fordelingen til de observerte dataene kalles den **empiriske** fordelingen. Den kan i mange tilfeller tilnærmes med en del standard, **teoretiske**, fordelinger.

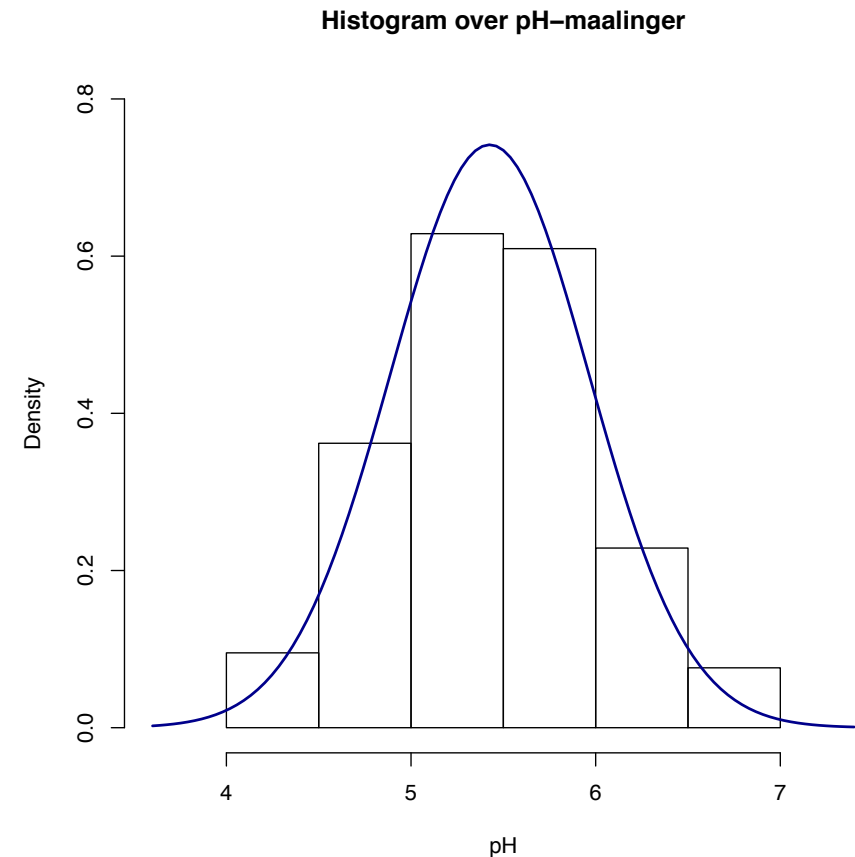
Vi vil nå legge til dette som et steg i det å utforske kvantitative data:

Utforske kvantitative data

1. Plott alltid dataene dine: lag en grafisk fremstilling.
2. Se etter det overordnede mønsteret (form, senter, og spredning) og slående avvik slik som uteliggere.
3. Beregn en numerisk oppsummering for å kortfattet beskrive senter og spredning
4. Noen ganger er det overordnede mønsteret av et stort antall av observasjonene så regulært at vi kan beskrive det med en glatt kurve.

Eksempel: Her er et histogram over pH-målinger i 105 prøver av regnvann.

Den glatte kurven tegnet inn sammen med histogrammet er en **matematisk modell** for fordelingen.

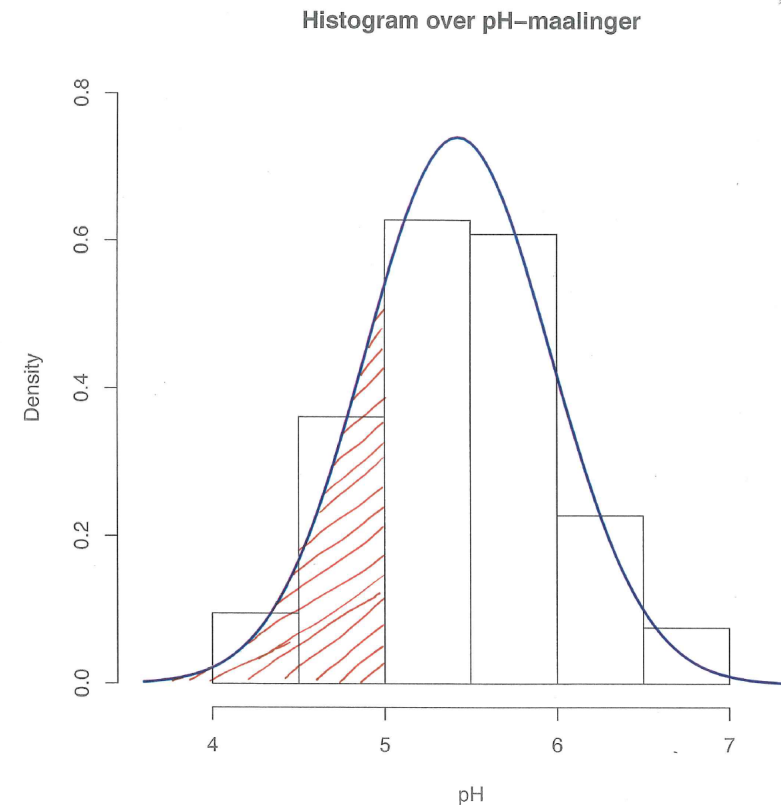


Tetthetskurver 2

51

De skyggelagte stolpene i dette histogrammet representerer andelen av pH-verdier i de observerte dataene som er mindre enn eller lik 5.0. Denne andelen er 0.229.

Her er arealet under den glatte kurven til venstre for 5.0 skyggelagt. Skalaen er justert slik at det totale arealet under hele kurven er 1 (derfor er denne kurven det vi kaller en **tetthetskurve**). Andelen av arealet under kurven som er til venstre for 5.0 er da lik 0.213.



En **tetthetskurve** er en kurve som

- er alltid på eller over den horisontale aksene.
- har et areal under seg som er eksakt lik 1.

En tetthetskurve er en **idealisering** som beskriver det overordnede mønsteret til en fordeling.

Arealet under kurven og over et bestemt område av verdier er lik andelen av alle observasjoner som faller i det området av verdier (i følge tetthetskurven).

Senter og spredning i tetthetskurver

53

Det norske begrepet **forventing** er for de **teoretiske tetthetskurvene** det **gjennomsnitt** er for **empiriske fordelinger** (observerte data). På engelsk brukes «mean» om begge deler. Begrepene median, kvartiler og standardavvik brukes for både tetthetskurver og observerte data.

- **Moden** til fordelingen er verdien der kurven er høyest
- **Kvartiler** (og dermed **median** og **IQR**) i en tetthetskurve kan omtrentlig lokaliseres visuelt ved å dele arealet under kuven i kvartdeler så godt man kan.
- Matematisk kan man finne arealer under kurver, og dermed median, kvartiler og IQR eksakt
- For en symmetrisk tetthetskurve er det lett å finne **forventning** og median visuelt
- For **skjeve fordelinger** kan **forventningen** være vanskelig å se visuelt, men finnes **matematisk**
- **Standardavviket** er vanskelig å se visuelt (bortsett fra for normalfordelingen), men kan beregnes **matematisk**

Tetthetskurver: Median og forventning

54

Forskjellen på medianen og forventningen til en tetthetskurve:

- **Medianen** til en tetthetskurve er “like-areal”-punktet —punktet som deler arealet under kurven i to.
- **Forventningen** til en tetthetskurve er balansepunktet, dvs punktet der kurven vill balansere dersom den var laget av solid materiale.
- Medianen og forventningen er like for en **symmetrisk tetthetskurve**. De ligger begge på midtpunktet til kurven.
- Forventningen til en **skjev kurve** er trukket bort fra medianen ut i den lange halen.

Teoretiske vs observerte størrelser

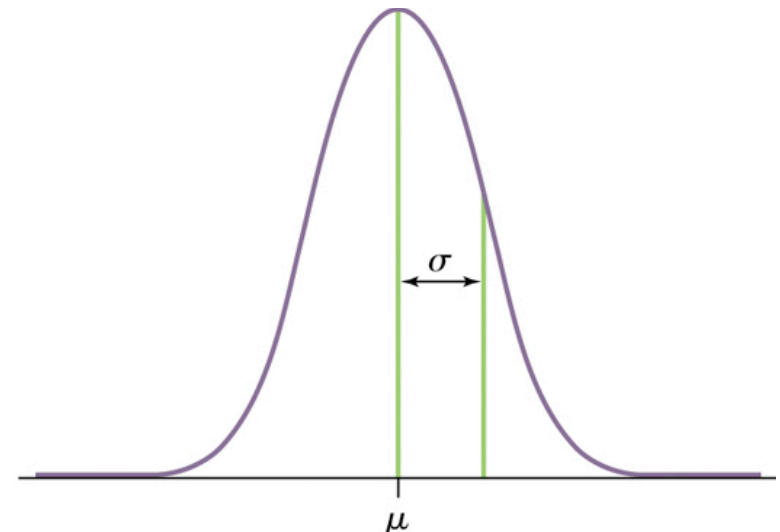
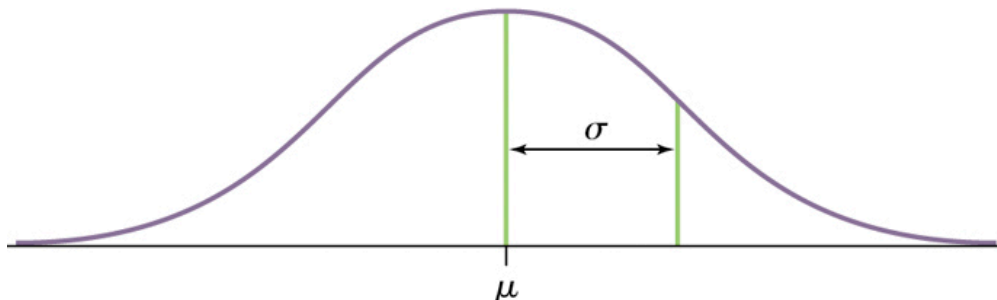
55

- Gjennomsnittet og standardavviket beregnet fra faktiske observasjoner (data) betegnes henholdsvis \bar{x} og s .
- Forventningen og standardavviket til den teoretiske fordelingen som tetthetskurven representerer betegnes henholdsvis μ (“mu”) og σ (“sigma”).
- En tetthetskurve er en **idealisert** beskrivelse av fordelingen til data, og det er **viktig å skille** mellom de **teoretiske størrelsene** forventning μ og standardavvik σ til en tetthetskurve og tallene \bar{x} og s som er **beregnet fra observerte data**

2 Kahoot-spørsmål

En spesielt viktig klasse av tetthetskurver er klassen av normalkurver, som beskriver normalfordelinger.

- Alle normalkurver er symmetriske, en-toppede, og klokke-formede.
- En spesifikk normalkurve beskrives fullstendig ved å angi forventning μ og standardavvik σ .
- Forventningen μ er midtpunktet
- Standardavviket σ er avstanden fra μ til der kurvaturen forandrer seg på begge sider

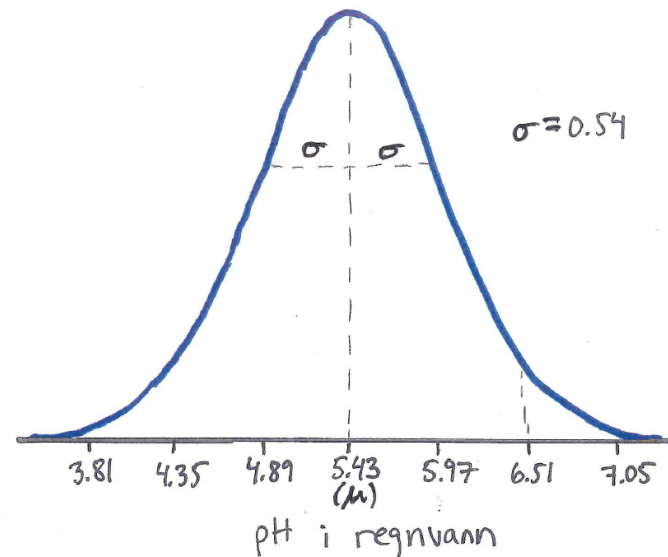


Normalfordelinger 2

- Vi forkorter normalfordelingen med forventning μ og standardavvik σ med $N(\mu, \sigma)$.
- Høyden til normalfordelingen med forventning μ og standardavvik σ for en gitt verdi av x er gitt av følgende funksjon

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

x kan f.eks. være pH i regnvann:



Normalfordelinger 3

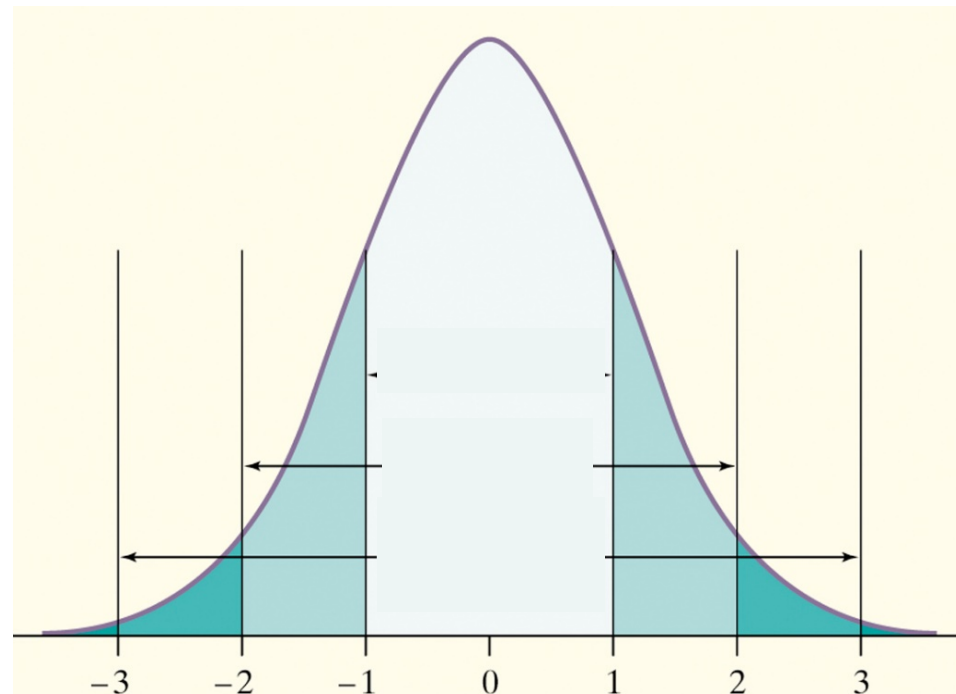
- Normalfordelinger er gode beskrivelser for mange fordelinger av reelle data, for eksempel egenskaper målt på mange individer
- Normalfordelinger er gode tilnærminger til mange former for tilfeldige utfall
- Mange statistiske metoder for å trekke slutninger fra data basert på normalfordelinger egner seg ofte også for andre tilnærmet symmetriske fordelinger
- Merk: Mange datasett følger ikke en normalfordeling!
Eksempel: Levetidsdata.

68-95-99.7-regelen 1

60

I normalfordelingen med forventning μ og standardavvik σ :

- tilnærmet **68%** av observasjonene faller innenfor σ fra μ .
- tilnærmet **95%** av observasjonene faller innenfor 2σ fra μ .
- tilnærmet **99.7%** av observasjonene faller innenfor 3σ fra μ .

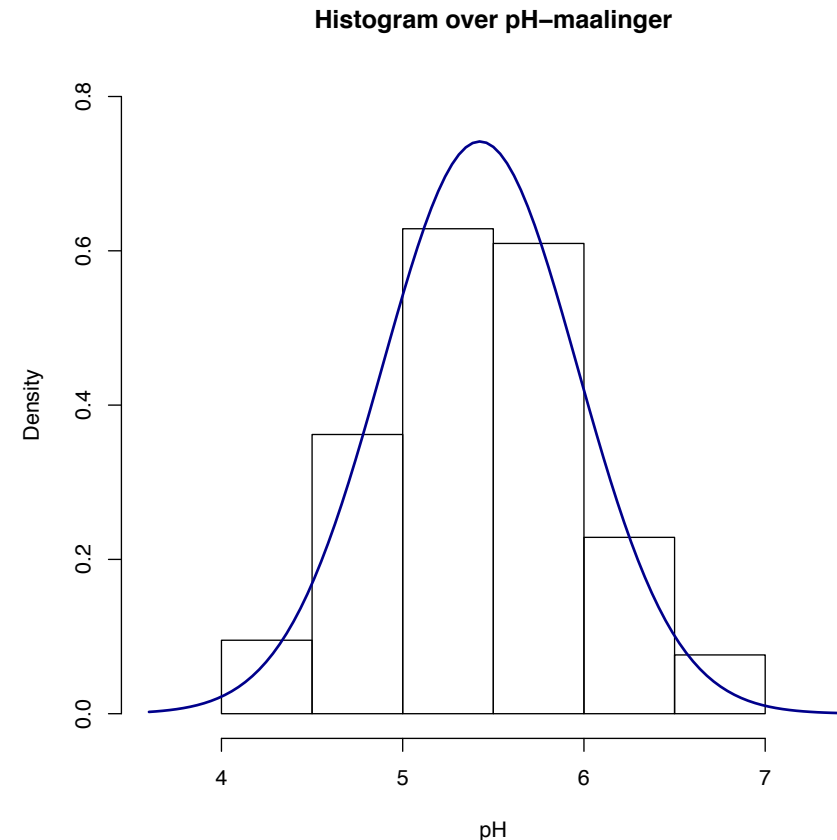


Normalfordeling: Eksempel

61

Eksempel pH: Histogram over pH-målinger i 105 prøver av regnvann.

Den glatte kurven tegnet inn sammen med histogrammet er en **normalfordelingskurve**. Det ser ut som om normalfordelingen er en god **tilnærming** til den observerte fordelingen.



68-95-99.7-regelen 2

Fordelingen til 105 pH-målinger er tilnærmet normal. Anta at fordelingen er $N(5.43, 0.54)$.

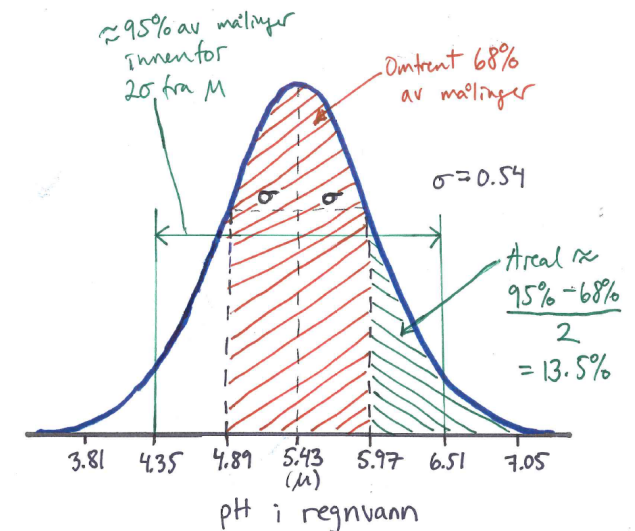
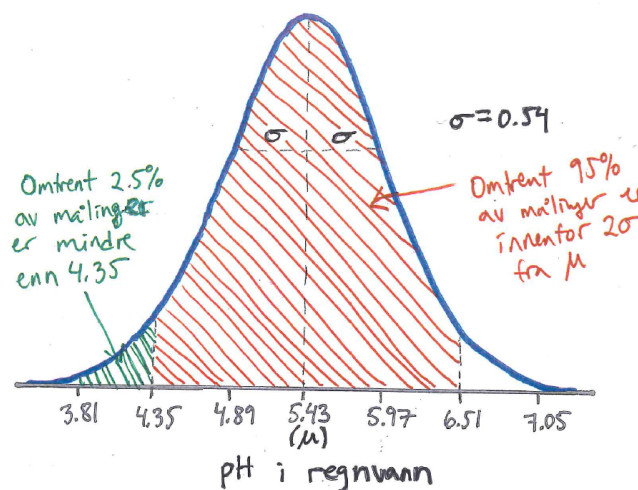
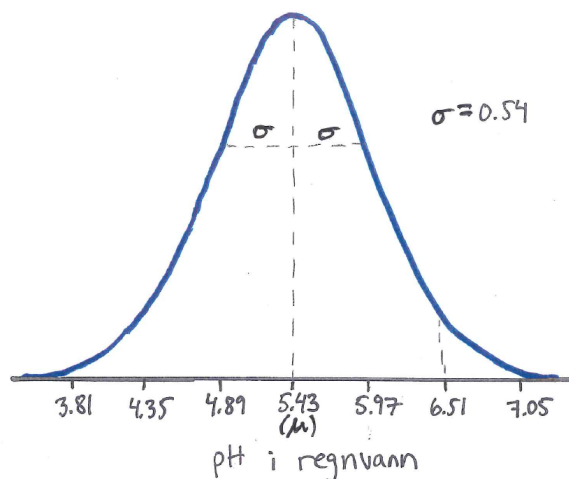
✓ Tegn den normale tetthets-kurven for denne fordelingen.

✓ Hvilken prosent av målinger er mindre enn 4.35?

Vi vet: ca 95% ligger i intervallet: $5.43 \pm 2 \times 0.54 = (4.35, 6.51)$

✓ Hvilken prosent av målinger er mellom 5.97 og 6.51?

Vi vet i tillegg: ca 68% ligger i intervallet: $5.43 \pm 0.54 = (4.89, 5.97)$



Standardisere observasjoner

63

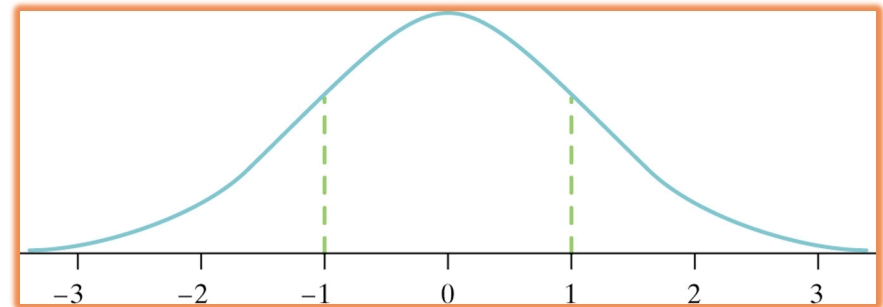
Hvis en variabel x har en fordeling med forventning μ og standardavvik σ , da er den **standardiserte verdien** til x (som også kalles **z-score**)

$$z = \frac{x - \mu}{\sigma}$$

Alle normalfordelinger er identiske hvis vi måler i enheter av størrelse σ fra forventningen μ .

Standard normal-fordelingen er normalfordelingen med forventning 0 og standardavvik 1, altså $N(0,1)$.

Dersom x har en fordeling med forventning μ og standardavvik σ , da er z-scoren standard normalfordelt.



Standard normal-tabellen 1

Fordi alle normalfordelinger er like når vi standardiserer, kan vi finne arealer under enhver normal-fordelings-kurve fra en enkelt tabell.

Standard-normal-tabellen

Tabell A i læreboken er en tabell med arealer under standard normal-kurven. Tabell-verdien for hver verdi z er arealet under kurven til venstre for z .

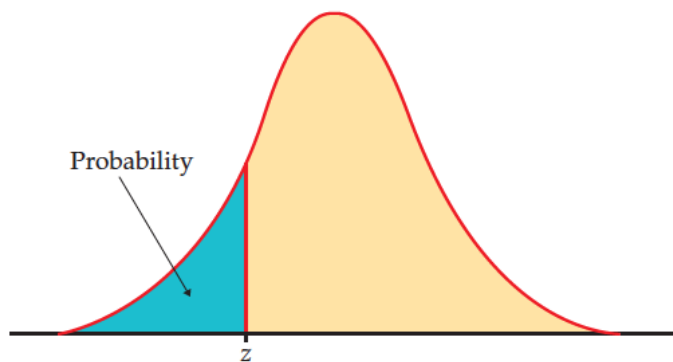
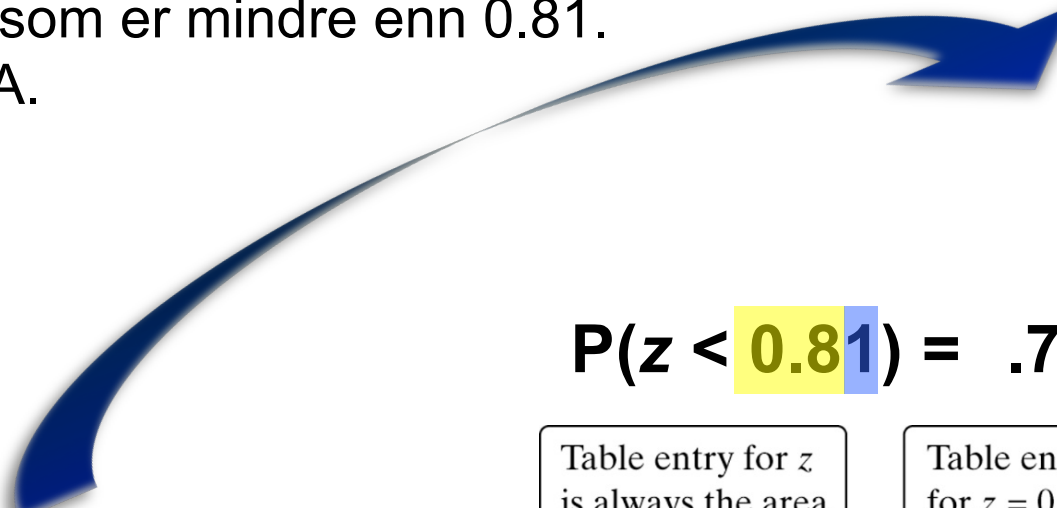


TABLE A Standard normal probabilities

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
-3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002
-3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
-3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
-3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
-3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
-2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
-2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
-2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
-2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
-2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
-2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
-2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
-2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
-2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
-2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
-1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
-1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
-1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
-1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
-1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
-1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681

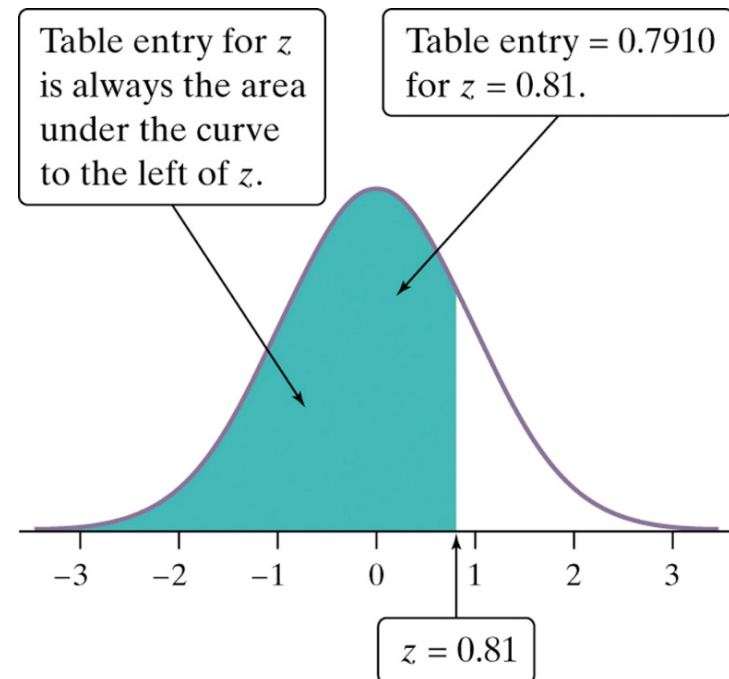
Standard normal-tabellen 2

Anta vi ønsker å finne andelen av observasjoner fra standard normal-fordelingen som er mindre enn 0.81.
Vi kan bruke tabell A.



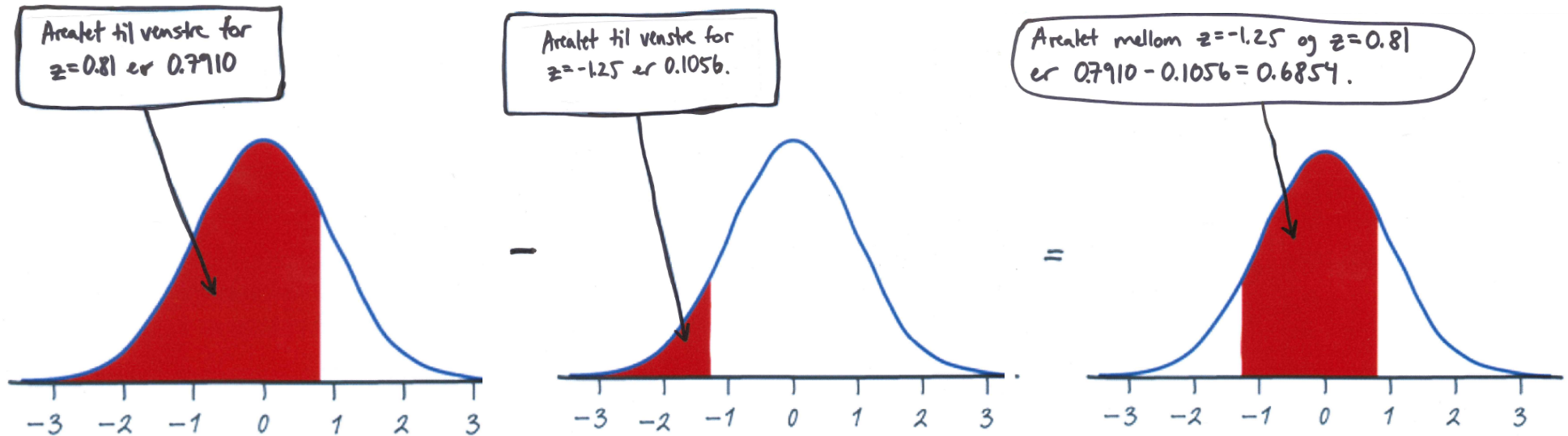
Z	.00	.01	.02
0.7	.7580	.7611	.7642
0.8	.7881	.7910	.7939
0.9	.8159	.8186	.8212

$$P(z < 0.81) = .7910$$

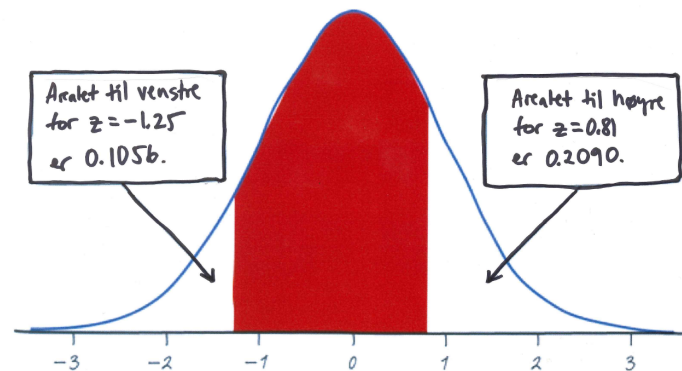


Normalberegninger 1

Finn andelen av observasjoner fra standard normalfordelingen som er mellom -1.25 og 0.81 .



Kan du finne den samme andelen ved å bruke en annen fremgangsmåte?
(Husk: Total areal under hele kurven er 1)



$$1 - (0.1056 + 0.2090) = 1 - 0.3146 \\ = \mathbf{0.6854}$$

Hvordan løse spørsmål som involverer normalfordelinger

Uttrykk spørsmålet i form av den observerte variabelen x .

Tegn et bilde av fordelingen og skravér arealet av interesse under kurven.

Utfør beregninger uten programvare:

- **Standardiser** x for å reformulere spørsmålet i form av en standard normal-variabel z .
- **Bruk Tabell A** og eventuelt det faktum at det totale arealet under kurven er 1 for å finne det ønskede arealet under standard normal-kurven.

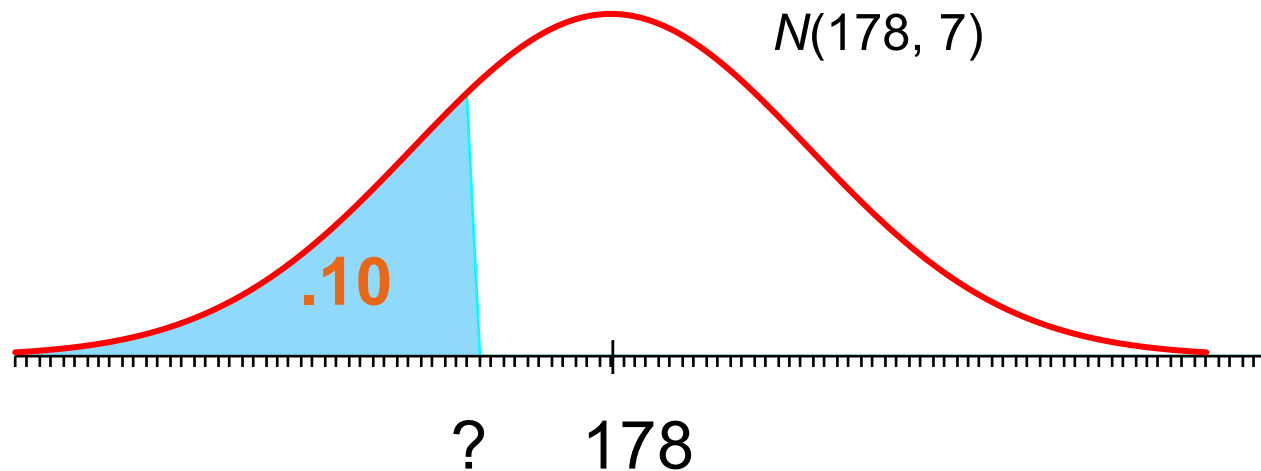
Skriv konklusjonen din i sammenhengen spørsmålet er formulert i.

Beregninger kan gjøres direkte ved bruk av programvare (f.eks. R), men for forståelsen er det nyttig å gjøre det «for hånd».

Normalberegninger 3

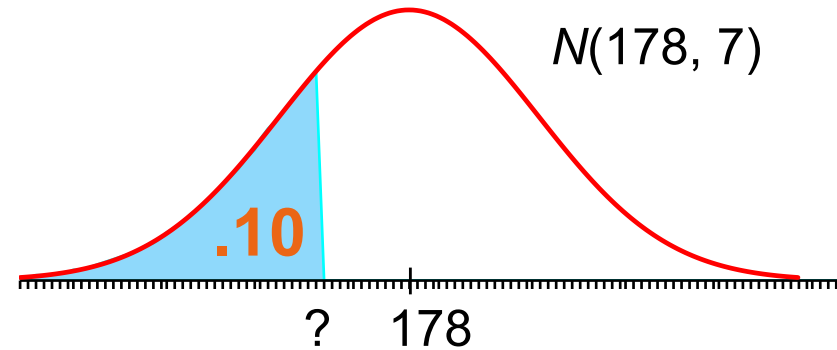
I følge en helse- og ernærings-studie fra 1976–1980 er høydene (i cm) til voksne menn i alderen 18–24 år $N(178, 7)$ -fordelte.

Hvis eksakt 10% av menn i alderen 18–24 er lavere enn en bestemt mann, hvor høy er denne mannen?



Normalberegninger 4

Hvor høy er en mann som er høyere enn eksakt 10% av men i alderen 18–24?



Finn sannsynligheten som er nærmest 0.10 i tabellen.

Finn den tilsvarende **standardiserte score**.

Verdien du er ute etter er så mange standardavvik fra forventningen.

z	.07	.08	.09
-1.3	.0853	.0823	.0823
-1.2	.1003	.0985	.0985
-1.1	.1210	.1190	.1170

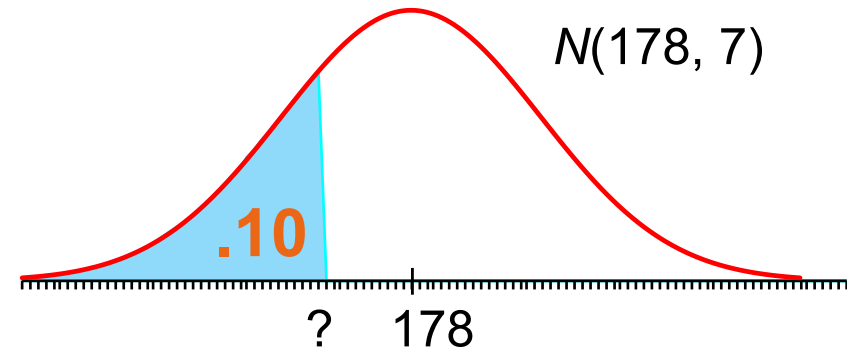
Z = -1.28

Normalberegninger 5

70

Hvor høy er en mann som er høyere enn eksakt 10% av menn i alderen 18–24?

$$Z = -1.28$$



Vi må reversere standardiseringen til z-scoren for å finne den tilsvarende høyden (x):

$$z = \frac{x - \mu}{\sigma} \implies x = \mu + z\sigma$$

$$\begin{aligned} x &= 178 + z \times 7 \\ &= 178 + [(-1.28) \times 7] \\ &= 178 + (-8.96) = \underline{\underline{169.04}} \end{aligned}$$

En mann måtte vært omtrent 169.04 cm høy eller lavere for å være blant de 10% laveste av alle menn i populasjonen.

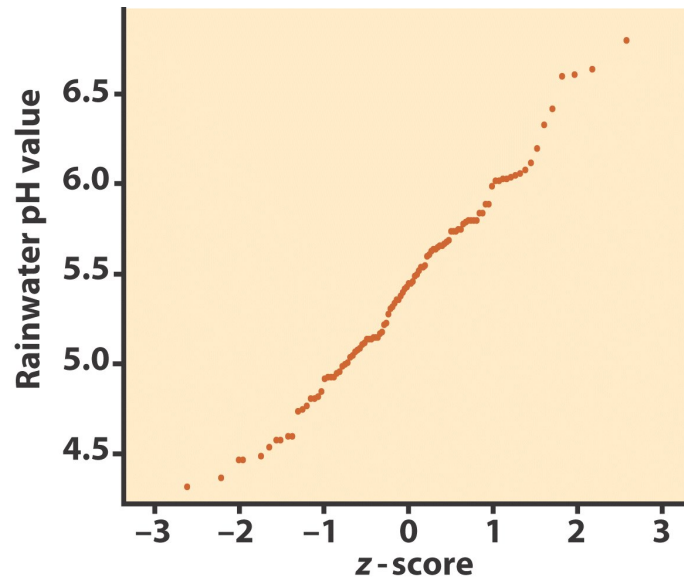
Histogrammer kan indikere om den empiriske fordelingen til de observerte dataene er tilnærmet normal, men en bedre måte å sjekke dette på er å plote dataene ved bruk av et **normalfordelingsplott**.

Fremgangsmåte normalfordelingsplott:

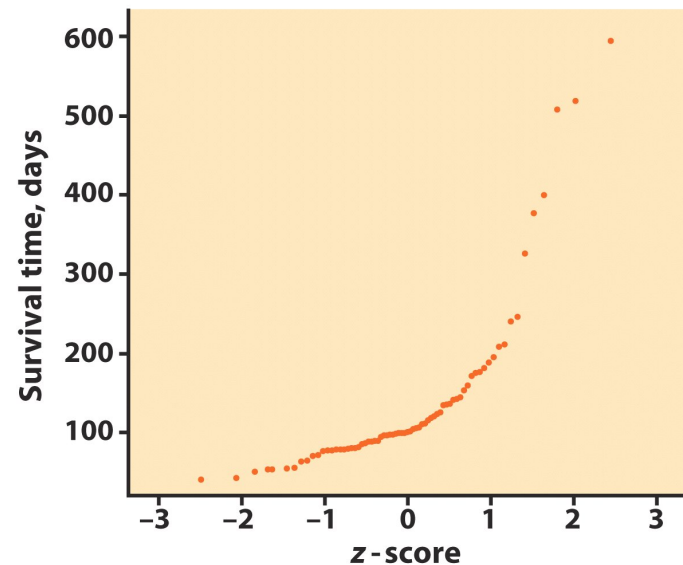
- Ranger observasjonene og finn hvilken persentil hver observasjon svarer til (hvor mange % av alle observasjonene er lik eller mindre enn denne observasjonen)
 - For hver av disse persentilene, finn z-scoren den svarer til ved å bruke Tabell A.
 - Plott verdien av hver observasjonen mot den tilhørende z-scoren.
- Plottet viser en rett linje: Indikerer godt samsvar mellom dataene og normalfordelingen, tilnærmet normal fordeling
- Systematiske avvik fra en rett linje: Indikerer en ikke-normal fordeling. Uteliggere sees som punkter som er langt fra det overordnede mønsteret i plottet.

Normalfordelingsplott 2

72



God tilnærming til en rett linje:
Fordelingen av målingene av
pH i regnvann er nær normal.



Krummet mønster: Dataene er ikke
normalfordelte. I stedet er dataene
høyre-skjeve: Noen individer har
spesielt lange levetider.

Normalfordelingsplott er arbeidskrevende å lage for hånd, men det finnes standard-funksjoner for å lage dem i R (og annen statistiske programvare).

Kapittel 1

Utforske data — Fordelinger:

Oppsummering

Introduksjon

1.1 Data

1.2 Presentere fordelinger med grafer

1.3 Beskrive fordelinger med tall

1.4 Tetthetskurver og normalfordelingen